The Small-World of Human Language

Ramon Ferrer i Cancho* and Ricard V. Solé*,+
* *Complex Systems Research Group, FEN-UPC, Campus Nord, B4-B5, Barcelona 08034 SPAIN*
+ *Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA*

Words in human language interact within sentences in non-random ways, and allow humans to construct an astronomic variety of sentences from a limited number of discrete units. This construction process is extremely fast and robust. The coocurrence of words within sentences reflect language organization in a subttle manner which can be described in terms of a graph of word interactions. Here we show that such graph displays two important features recently found in a disparate number of complex systems: (a) The so called small world effect. In particular, the average distance between two words d (i.e. the average minimum number of jumps to be made from an arbitrary word to another) is shown to be $d \approx 2-3$, in spite that the human brain can store many thousands. (b) A scale-free distribution of degrees. The known dramatic effects of disconnecting the most connected vertices in such networks can be identified in some language disorders. These observations suggest some unexpected features of language organization that might reflect the evolutionary and social history of lexicons and the origins of their flexibility and combinatorial nature.

Keywords: Small-world, Scaling, Lexical networks, Human language

I. INTRODUCTION

The emergence of human language is one of the major transitions in evolution (Smith & Szäthmáry, 1997). Living humans posses a unique symbolic mind capable of language which is not shared by any other species. Over two million years of hominid evolution, a coevolutionary exchange between languages and brains took place (Deacon, 1997). This process involved the (possible sudden) transition from non-syntactic to syntactic communication (Nowak & Krakauer, 1999; Nowak et al., 2000). Human language allows the construction of a virtually infinite range of combinations from a limited set of basic units. The process of sentence generation is astonishingly rapid and robust and indicates that we are able to rapidly gather words to form sentences in a highly reliable fashion.

A complete theory of language requires a theoretical understanding of its implicit statistical regularities. The best known of them is the Zipf's law, which stays that the frequency of words decays as a power function of its rank (Zipf, 1972). However, in spite of its relevance and universality (Balasubrahmanyan & Naranan, 1996), such law can be obtained from a variety of mechanisms (Nicolis, 1991; Simon, 1955; Li, 1992) and does not provide deep insight about the organization of language. The reason is that information transmission is organized into sentences, made by words *in interaction*.

Human brains store lexicons usually formed by thousands of words. Estimates are in the range $10^4 - 10^5$ words (Romaine, 1992; Miller & Gildea, 1987). Besides, the contents of the lexicon of individuals of the same language vary depending on many factors such as age, geographic location, social context, education and profession.

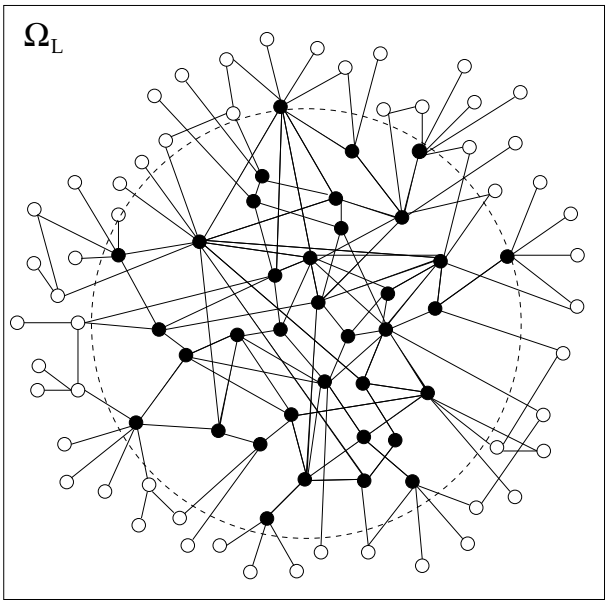

FIG. 1. A possible pattern of wiring in Ω_L. Black nodes are common words and white nodes are rare words. Two words are linked if they cooccur significantly.

Being the primary goal of a lexicon to achieve a sucessful communication, there must exist a common lexicon for sucessful basic communication between speakers, hereafter named a kernel lexicon, to surmount the limitations imposed by the factors mentioned above. Obviously, the better candidates to form this lexicon are the

most frequent words. Actually, the analysis of multi-speaker corpora shows two different regimes dividing words into basic and specialized words (Ferrer & Solé, 2000).

Words interact in many ways. Some words cooccur with certain words with higher probability than with others and coocurrence is not of trivial nature, i.e. it is not a straigtforward implication of the known frequency distribution of words. If a text is scrambled the frequency distribution is mantained but its content will not make sense.

In this paper we show that the coocurrence of words in sentences relies on the network structure of the lexicon whose properties are analyzed in depth. As we will show in this paper, human language can be described in terms of a graph of word interactions. This graph has some unexpected properties (shared by other biologic and technologic networks (Strogatz, 2001)) that might underly its diversity and flexibility and open new questions about its origins and organization.

II. GRAPH PROPERTIES OF HUMAN LANGUAGE

Words cooccur in sentences. Many coocurrences are due to syntactical relationships between words, e.g. head-modifier or dependency relationships (Melčuck, 1989). Some others are due to stereotiped expressions or collocations that work together (e.g. *take it esasy, New York*). We will define links as significative cooccurrences between words. We do not seek to provide a detailed list of the origins and linguistic interpretation of such significative cooccurrences but in showing that they exist and can be captured using quantitative measures of correlation regardless of their nature. A first approach for estimating the network of the lexicon is to consider that there is a link between every pair of neighbouring words (at the risk of capturing spurious correlations).

Let us consider the graph of human language, Ω_L, as defined by $\Omega_L = (W_L, E_L)$, where $W_L = \{w_i\}, (i = 1, ..., N_L)$ is the set of N_L words and $E_L = \{\{w_i, w_j\}\}$ is the set of edges/connections between words. Here $\xi_{ij} = \{w_i, w_j\}$ indicates that there is an edge (and thus a link) between words w_i and w_j. Two connected words are *adjacent* and the *degree* of a given word is the number of edges that connect it with other words. Figure 1 shows how such a network would look like.

Recent research on a number biological, social and technological graphs revealed that they share a common feature: the so called small world (SW) property (Watts & Strogatz, 1998; Watts, 1999; Newman, 2000). Small world graphs have a number of surprising properties that make them specially relevant to understand how interactions among individuals, metabolites or species lead to the robustness and homeostasis observed in nature (Watts & Strogatz, 1998). The SW pattern can be detected from the analysis of two basic statistical properties: the so called *clustering coefficient* C_v and the *path length* d. Let us consider the set of links ξ_{ij} $(i, j = 1, ..., N_L)$, where $\xi_{ij} = 1$ if a link exists and zero otherwise and that the average number of links per word is $\bar{k}$. Let us indicate by $\Gamma_i = \{s_i | \xi_{ij} = 1\}$ the set of nearest neighbors of a word $w_i \in W_L$. The clustering coefficient for this word is defined as the number of connections between the words $w_j \in \Gamma_i$. By defining

$$\mathcal{L}_i = \sum_{j=1}^{N_L} \xi_{ij} \left[\sum_{k \in \Gamma_i ; j<k} \xi_{jk} \right] \tag{1}$$

we have:

$$c_v(i) = \frac{\mathcal{L}_i}{\binom{|\Gamma_i|}{2}}$$

so that the clustering coefficient is the average over W_L:

$$C_v = \frac{1}{N_L} \sum_{i=1}^{N_L} c_v(i) \tag{2}$$

and measures the average fraction of pairs of neighbors of a node that are also neighbors of each other.

The second measure is easily defined. Given two words $w_i, w_j \in W_L$, let $d_{min}(i, j)$ the minimum path length connecting these two words in Ω_L. The average path length of a word will be defined as

$$d_v(i) = \frac{1}{N_L} \sum_{j=1}^{N_L} d_{min}(i, j) \tag{3}$$

and thus the average path length d will be:

$$d = \frac{1}{N_L} \sum_{i=1}^{N_L} d_v(i) \tag{4}$$

Graphs with Small World structure are highly clustered but d will be small. Random graphs (where nodes are randomly wired) are not clustered and have also short d (Watts, 1999). At the other extreme, regular lattices with only nearest-neighbor connections among units, are typically clustered and exhibit long paths. It has been shown, however, that a regular lattice can be transformed into a SW if a small fraction of nodes are rewired to randomly chosen nodes. Thus a small degree of disorder generates short paths (as in the random case) but retaining the regular pattern (Watts and Strogatz, 1998).

For random graphs that $C_v^{rand} \approx \bar{k}/N$. For SW graphs, d is close to the one expected from random graphs, d^{rand}, with the same $\bar{k}$ and $C_v \gg C_v^{rand}$. These two conditions are taken as the standard definition of SW. SW graphs have been shown to be present in both social and biological networks (Jeong et al., 2000; Montoya & Solé,

2000; Solé & Montoya, 2000; Newman, 2000; Strogatz, 2001). Besides, some of these networks also exhibit scaling in their degree distribution. In other words, the probability $P(k)$ of having a node with degree k scales as $P(k) \approx k^{-\gamma}$. We have found that the graph of human language displays similar properties. This second property has been shown to be related with an extremely high stability against perturbations directed to randomly chosen nodes and a high fragility when perturbations are directed to highly connected ones (Albert et al., 2000). As we will show here, Ω_L exhibits both SW structure and a power laws in $P(k)$.

III. LINK COLLECTION

The most correlated words in a sentence are the closest. A decision must be taken about the maximum distance considered for forming links. If the distance is long, the risk of capturing spurious coocurrences increases. If the distance is too short, certain strong coocurrences can be sistematically not taken into account. We decided the maximum distance according to the minimum distance at which most of the coocurrences are likely to happen:

- Many coocurrences take place at distance 1, e.g. red flowers (adjective-noun), the/this house (article/determiner-noun), stay here (verb-adverb), getting dark (verb-adjective), can see (modal-verb), ...
- Many coocurrences take place at distance 2, e.g. hit the ball (verb object), Mary usually cries (subject-verb), table of wood (noun-noun through a prepositional phrase), live in Boston (verb-noun), ...

Long distance correlations, i.e. at distance greater than two, have been shown to take place in human sentences (Chomsky, 1957). Here we stop our seek at distance two. The reason is fourfold:

- Considering whatever distance requires an automatic procedure for acomplishing the task of capturing the relevant links. We do not know of any computational technique that succesfully performs this task for a general case. From the practical point of view, a context of two words is considered to be the lowest distance at which most of the improvement of disambiguation methods is achieved (Kaplan, 1955; Choueka & Lusignan, 1985).
- Our method fails to capture the exact relations happening in a particular sentence but captures (almost) every possible type of links. The type of the link is determined by the syntactic categories/roles of the intervening words. Very few types of links (if any) are observed at distance greater than 2 and not at lower distances.
- We are not interested in all the relations happening in a particular sentence. Our goal is to capture as much links as possible through an automatic procedure. If the corpus is big enough, the macroscopic properties of the network should emerge.
- Being syntactic dependencies non-crossing (Hudson, 1984; Melčuck, 1989), a long distance syntactic link implies the existence of lower distance syntactic links. In contrast, a short distance link do not imply a long-distance link.

The technique can be improved by choosing only the pairs of consecutive words whose mutual cooccurrence is larger than the one expected from their chance. This can be measured with the condition $p_{ij} > p_i p_j$ which defines the presence of real correlations beyond the expected from a random ordering of words. If a pair of words cooccurs less times than what it would be expected when independence between such words is assumed, the pair is considered to be spurious. Graphs in which this condition is used will be called restricted (unrestricted otherwise).

IV. SCALING AND SW PATTERNS

The networks resulting from the basic and improved methods will be called, respectively, the unrestricted word network (UWN) and the restricted word network (RWN). They have $N(UWN) = 478,773$ and $N(RWN) = 460,902$ nodes with $E(UWN) = 1.77 \cdot 10^7$ and $E(RWN) = 1.61 \cdot 10^7$ edges, respectively. With average connectivities of $\bar{k}_{uwn} = 74,2$ and $\bar{k}_{rwn} = 70,13$, their clustering and path lengths are indicated in Table 1 .

Figure 2 shows the distribution of degrees of both the UWN and RWN obtained after processing about 3/4 of the million words of the British National Corpus (about 70 million words). The obvious limitations of our methods are overcome by the use of a big amount of data. The distribution of connectivies of UWN and RWN decays with two different average exponents each, $\gamma_1 = -1.50$ for the first regime and $\gamma_2 = -2.70$ for the second regime, respectively. The exponent in the second regime is similar to that of the so-called Barabási-Albert (BA) model ($\gamma_{BA} = -3$) (Barabási & Albert, 1999). The BA model is an independent rediscovery of earlier work by Herbert Simon on systems with skewed distributions (Simon, 1955). Using the rule of preferential attachment, they showed that scale-free distributions are obtained. The rule simply assumes that new nodes in the growing network are preferentially attached to an existing node with a probability proportional to the degree of such node.

Furthermore, word networks have small-word features. The average minimum distance between vertices is below

3 (2.63 for the UWN and 2.67 for the RWN), so reaching whatever vertex involves less than three jumps on average. This is significantly important, since the network contains about $4.7 \cdot 10^5$ different words. Clustering (0.687 for the UWN and 0.437 for the RWN) is far from the randomness expectation ($1.55 \cdot 10^{-4}$ for both the UWN and the RWN) in both cases.

As far as we know, this is the first time that such a statistically significant property has been reported about the organization of human language. In spite of the huge amount of words that can be stored by a given human, whatever word in the lexicon can be reached *with less than three* intermediate words on average. If a word is reached during communication, jumping to another word requires very few steps. Speed during speech production is important and can be more easily achieved if intervening words are close each other in the underlying structure used for the construcion of sentences. On the other hand, richness is another quality of a powerful communication. Although words are preferably choosen from the kernel lexicon, external words are at a short distance, so rich communication based on the word network can be attained with little increase in effort.

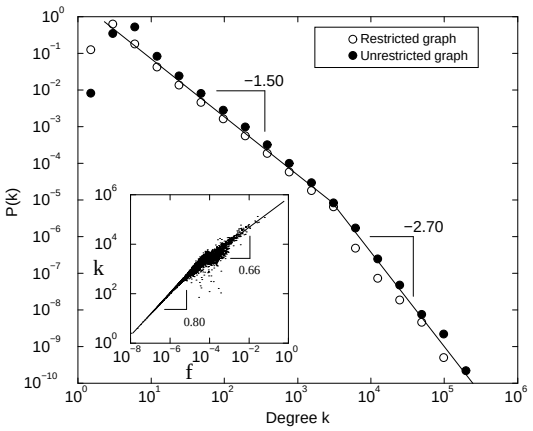

FIG. 2. Connectivity distribution for the unrestricted word network (circles) and the restricted word network (squares). Points are grouped by powers of two. Inset: average degree as a function of frequency. Degree increases as a function with frequency with exponent 0.80 for the first domain and 0.66 for the second one.

It is well known that the more frequent a word, the more available it is for production (Brown & McNeil, 1966; Kempen & Huijbers, 1983) and comprehension (Forster & Chambers, 1973; Scarborough et al., 1977) processes. This phenomenon is known as the *frequency* (referring to the whole individual's experience) or *recency* (referring to the recent individual's experience) *effect* (Akmajian et al., 1995). This phenomenon will serve us to show that preferential attachment is very likely to be shaping the scale-free distribution of degrees in a way similar to the BA model. For the most frequent words,

$$k \propto f^{0.66}$$

where k is the degree and f is the frequency of the word. We can then recast the *frequency effect* in terms of the degree as *the higher the degree of a word, the higher its availability*. In other words, links including highly connected words are *preferred*. Inset in Figure 2 shows the complete relationship between f and k in RWN.

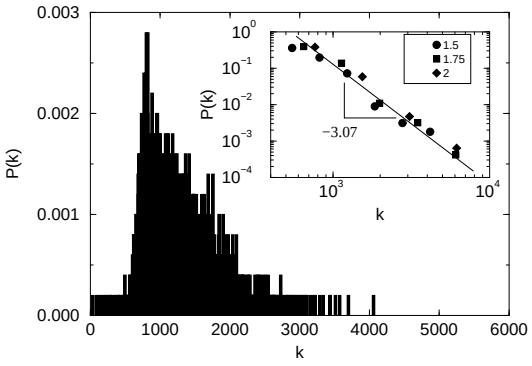

FIG. 3. Connectivity distribution for the kernel word network, formed by the 5000 most connected vertices in RWN. Inset: power law tail for $k > k$ calculated by grouping in powers of 1.5, 1.75 and 2. The exponent of the power tail is $\gamma_{KWN} \approx -3$, suggesting that preferential attachment is at play.

The exponent of UWN and RWN is closer to $\gamma_{BA} = -3$ in the second regime of the distribution which is where the frequency effect makes much more sense. The kernel lexicon contains words common to the whole community of speakers. Beyond the kernel, a certain word is unknown for one speaker and familiar for the another. The recency effect then cannot be applied for all the individuals contributing to shape the underlying lexicon network. It is thus expected that the network formed exclusively by the interaction of kernel words, hereafter referred as the kernel word network (KWN), better agrees with the predictions that can be performed when preferential attachment is at play. Figure 3 shows the log-normal appearance of the connectivity distribution. The power tail has exponent $\gamma_{KWN} \approx -3$, consistent with the Barabási-Albert model (Barabási & Albert, 1999) and the differences respect to it require special attention. It is important to notice that the kernel lexicon is a versatile subset of the repertoire of individual speakers. A few thousand words must be able to say everything or almost everything. Even when lexicons become very small, i.e. pidgin languages whose lexicons do not usually exceed about 1,000 words (Romaine, 1992), it has been pointed out they allow to say everything that can be said in a complex lexicon (e.g. English) at the expense of high redundancy (recurring to circumlocution). The average

connectivity in the kernel is $\bar{k} = 1219$. A first consequence is that words with low connectivity must be rare. Having rather useless words in this critical subset is an enormous waste. Once connected words become frequent in the distribution, the network organizes in a scale-free way. We believe that the scale-freeness is responsible for the ability-to-say-everything of the kernel. A non-trivial network is needed since every word on average is connected to 24% of the rest of the kernel words.

V. DISCUSSION

We have shown that the graph connecting words in language exhibits the same statistical features than other complex networks. The short distance between words arising from the SW structure strongly suggests that language evolution might have involved the selection of a given graph of connections between words. Future work should address this problem theoretically, perhaps using an evolutionary language game model (Nowak & Krakauer, 1999; Nowak et al., 2000) where a pay-off associated to the graph structure is introduced. Concerning the scaling in $P(k)$ and the observed exponents, this pattern also calls for an evolutionary explanation. The word network is the result of a growth process in which new words are added and are likely to be linked to highly connected existing words.

If the small-world features derive from optimal navigation needs, two predictions can be formulated. First, the existence of words whose main purpose is to speed-up navigation. Second, deriving from the first, the existence of brain disorders characterized by navigation deficits in which such words are involved. The best candidates for answering the first question are the so-called particles, a subset of the function words (e.g. articles, prepositions, conjunctions) formed by the most frequent among them (e.g. *ant*, *the*, *of*, ...)[1]. These words are characterized by a very low or zero semantic content. Although they are supposed to contribute to the sentence structure, they are not generally crucial for sentence understanding. A compelling test of this statement is that particles are the first words to be suppressed in telegraphic speech (Akmajian et al., 1995).

The answer to the second prediction is agrammatism, a kind of aphasia in which speech is nonfluent, labored, halting and lacking of function words (and thus of particles). Agrammatism is the only syndrome in which function words are particularly ommited (Caplan, 1987). Function words are the most connected ones. We suggest that such halts and lack of fluency are due to fragility associated to removal of highly connected words. Although scale-free networks are very tolerant to random removal of vertices, if deletion is directed to the most connected vertices the network gets broken into pieces (Albert et al., 2000).

It is known that function words omission is often accompanied by substitutions of such words. Patients in which substitutions predominate and speach is fluent are said to undergo paragrammatism (Caplan, 1994). We suggest that paragramatism recovers fluency (i.e. low average word-word distance) by unapropriately using the remaining highly connected vertices and thus often producing substitutions of words during discourse.

[1] Accordint to our calculations, the 10 most connected words are *and*,*the*, *of*, *in*, *a*, *to*, *'s*, *with*, *by* and *is*.

ACKNOWLEDGMENTS

Authors want to specially thank G. Miller and F. Diéguez for valuable discussions and are also grateful to D. Krakauer, A. Lloyd, M. Nowak, F. Reina and J. Roselló for helpful comments. RFC acknowledges the hospitality of the Institute for Advanced Study. This work was supported by the Santa Fe Institute (RVS) and grants of the Generalitat de Catalunya (FI/2000-00393, RFC) and the CICYT (PB97-0693, RVS).

References

Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (1995). *Linguistics. an introduction to language and communication.* MIT Press. (Chapter 2)

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, *406*, 378-381.

Balasubrahmanyan, V. K., & Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics*, *3*(3), 177-228.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509-511.

Brown, R., & McNeil, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behaviour*, *5*, 325-337.

Caplan, D. (1987). *Neurolinguistics and linguistic aphasiology.* Cambridge University Press.

Caplan, D. (1994). *Language. structure, processing and disorders.* The MIT Press.

Chomsky, N. (1957). *Syntactic structures.* Mouton.

Choueka, Y., & Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, *19*, 147-157.

Deacon, T. W. (1997). *The symbolic species: the co-evolution of language and the brain.* W. W. Norton & Company.

Ferrer, R., & Solé, R. V. (2000). Two regimes in the frequency of words and the origin of complex lexicons. *Santa Fe Working Paper 00-12-068.* (Submited to the Journal of Quantitative Linguistics)

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behaviour*, *12*, 627-635.

Hudson, R. (1984). *Word grammar.* B. Blackwell.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & A.-L.Barabási. (2000). The large-scale organization of metabolic networks. *Nature*, *407*, 651-654.

Kaplan, A. (1955). An experimental study of ambiguity and context. *Mechanical Translation*, *2*, 39-46.

Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, *14*, 185-209.

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, *38*(6).

Melčuck, I. (1989). *Dependencies grammar: Theory and practice.* New York, University of New York.

Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American*, *257*(3).

Montoya, J. M., & Solé, R. V. (2000). Small world patterns in food webs. *Santa Fe Working Paper 00-10-059.* (Submitted to J. Theor. Biol.)

Newman, M. E. J. (2000). Models of the small-world. *Journal of Statistical Physics*, *101*(3/4), 819-841.

Nicolis, J. S. (1991). *Chaos and information processing.* Singapore: World Scientific.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proc. Natl. Acd. Sci. USA*, *96*, 8028-8033.

Nowak, M. A., Plotkin, J. B., & Jansen, V. A. (2000). The evolution of syntactic communication. *Nature*, *404*, 495-498.

Romaine, S. (1992). The evolution of linguistic complexity in pidgin and creole languages. In J. A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages* (p. 213-238). Addison Wesley.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1-17.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*, 425-440.

Smith, J. M., & Száthmáry, E. (1997). *The major transitions in evolution.* Oxford University Press.

Solé, R. V., & Montoya, J. M. (2000). Complexity and fragility in ecological networks. *Santa Fe Working Paper 00-11-060.*

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, *410*, 268-276.

Watts, D. J. (1999). *Small-worlds.* Princeton University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440-442.

Zipf, G. K. (1972). *Human behaviour and the principle of least effort. an introduction to human ecology.* New York: Hafner reprint. (1st edition: Cambridge, MA: Addison-Wesley, 1949)

graph	C	C_{random}	d	d_{random}
Ω_L (UWN)	0.687	$1.55 \cdot 10^{-4}$	2.63^*	3.03
Ω_L (RWN)	0.437	$1.55 \cdot 10^{-4}$	2.67^*	3.06

TABLE I. Word network patterns. It can be seen that $C \gg C_{random}$ and $d \lesssim d_{random}$, consistently with a SW network. All values are exact except those marked with *, which have been estimated on a random subset of the vertices.